국제사회보장리뷰 2021 가을호 Vol. 18, pp. 133~138

미국 아동복지 분야에서의 기계학습 활용과 공정성 이슈

Machine Learning in Child Welfare: Application and Algorithmic Fairness

안은혜(서던캘리포니아대학교 박사과정)
Ahn, Eunhye(University of Southern California, PhD Candidate)

1. 들어가며

매년 미국에서는 약 700만 명의 아동이 학대와 방치로 신고된다. 이 아동 및 가족들을 지원하기 위해 아동복지 전문가들은 매일 수많은 중요한 결정을 내려야 한다(U. S. Department of Health & Human Services, 2021). 아동 개인뿐 아니라 아동의 가족 구성원 전체에 중요한 영향을 미칠 수 있는 의사 결정 절차의 정확도를 높이고 안정성을 확보하기 위해 미국 아동복지 분야에서는 지난 수십 년간 데이터를 활용하여 의사 결정 절차를 개선하는 방안들이 다양하게 연구되어 왔다(Brownell & Jutte, 2013; English et al., 2000). 최근에는 공공 분야 데이터의 발달과 데이터 분석 기술의 진보로 아동복지에서도 공공 데이터(administrative data)와 기계학습(machine learning)을 활용한 의사 결정 시스템에 대한 논의가 시작되었다. 이 가운데 기계학습 알고리즘에 기반한 위험예측모델(predictive risk modeling)을 의사 결정 보조도구로 활용할 수 있는 가능성이 주목받고 있다(Ahn et al., 2021; Chadwick Center and Chapin Hall, 2018; Schwartz et al., 2017). 이와 함께 알고리즘을 활용한 의사 결정 모델들이 초래할 수 있는 윤리적 이슈들에 대한 관심이 증가했고, 그중에서도 의사 결정 모델의 공정성에 대한 관심이 커지고 있다(Dare, 2015). 최근 미국 아동복지에서 주목받는 기계학습 활용 방향 및 윤리와 공정성 이슈를 살펴보고, 이와 관련하여 기계학습 학자들이 제안하는 공정성(machine learning fairness) 개

념의 적용 여부 및 한계를 논의하고자 한다.

2. 아동복지 시스템 내 기계학습 활용

기계학습은 알고리즘적 접근 방식을 사용하여 주어진 데이터 내 예측 변수와 결과 변수 간의 관계를 찾아 학습함으로써 우선순위 지정, 분류, 연결 및 필터링을 포함해 다양한 의사 결정의 효율성과 정확성을 향상시킨다(Hastie et al., 2009). 지난 수십 년 동안 기계학습은 의료 및 사법 시스템을 포함해 다양한 분야에서 의사 결정을 최적화하기 위해 적극 사용되었으며(Berk & Bleich, 2013; Ustun & Rudin, 2016), 아동복지 분야에서 기계학습은 주로 위험예측모델 개발을 위해 활용되었다. 위험예측모델은 과거 데이터를 이용하여 아동 및 가족에게 닥칠 수 있는 미래의 위험을 분석하고 점수화하는 데 사용된다. 예를 들어 이전 연구들에서는 기계학습을 활용한 아동 위탁보호(foster care placement)(Chouldechova et al., 2018), 아동학대 재신고(Schwartz et al., 2017; Shroff, 2017), 학대로 인한 부상(Vaithianathan et al., 2013), 무연고 청소년의 보호 종료(aging out without permanency)(Ahn et al., 2021), 청소년의 만성 노숙인 경험(experiencing chronic homelessness among transitional age youth)(Chan et al., 2017), 고위험 출산(Pan et al., 2017) 등과 관련된 위험을 예측하려는 시도를 하였다. 일부 연구들은 기계학습의 텍스트 마이닝(text mining) 알고리즘을 활용하여 인간의 자연어 기록에 나타나는 패턴을 인식하고 분석함으로써 아동복지 데이터 및 의료 기록에 나타난 아동학대를 식별하는 가능성을 보여 줬다(Amrit et al., 2017; Perron et al., 2019).

3. 아동복지 내 기계학습 활용과 공정성 문제

기계학습이 아동보호 및 복지 서비스에 제공하는 접근 방식에 대한 관심이 높아짐에 따라 기계학습 사용 윤리에 대한 논의도 증가하고 있다(Brown et al., 2019; Dare, 2015; Keddell, 2015). 기계학습 사용 윤리에 관한 논의들은 예측 모델의 효율성, 알고리즘적 책임(algorithmic accountability), 투명성 및 해석 가능성(transparency and interpretability), 데이터 프라이버시(data privacy), 알고리즘에 기반한 서비스 제공이 아동과 가족들에게 미치는 직간접적 영향을 포함한다. 최근 연구에 따르면 데이터 수집, 모델 개발, 모델 적용 및 활용에 이르는 모든 단계에서 윤리적 문제가 발생할 수 있다(Barocas & Selbst, 2016; Dare, 2015; Mehrabi et al., 2019). 이런 문제들은 아동복지의 고유한 특성들(예를 들면 아동 및 가족의 취약성,

경제적 불균형, 서비스 이용과 관련한 사회적 낙인, 아동과 보호자의 이해 상충 등)과 결합되어 더 복잡한 문제를 만들기도 한다.

다양한 윤리적 화두 중 하나인 공정성 문제는 알고리즘 모델이 어떻게 인종과 같은 개인적 특성에 따라 아동과 가족들을 차별하는지, 그리고 이것이 어떻게 사회의 불평등을 지속·심화시킬 수 있는지에 초점을 맞추고 있다. 아동복지 시스템에서 기계학습의 공정성 이슈가 특히 관심을 받는 이유는 알고리즘에 기반한 의사 결정 모델이 기존 데이터 내의 사회적, 인종적, 경제적 편향을 학습함으로써 현재 사회의 인종적, 사회적 불평등을 지속·심화시킬 수 있기 때문이다(Barocas & Selbst, 2016). 또한 기계학습은 데이터에 나타난 변수 간 특성들을 학습할 때 소규모 그룹보다는 큰 규모의 그룹을 우선시하기 때문에 아동복지 시스템이 주목하는 상대적으로 적은 수의 소외 계층 아동 및 그 가족들의 경제적, 사회적 취약성을 악화시킬 수 있다(Capatosto, 2017). 미래를 예측하는 시스템이라면 필연적으로 예측 오류가 발생하는데, 취약한 아동일수록 이런 오류에 더 치명적인 영향을 받게 된다. 설사 예측이 오류가 아니더라도 위험에 처한 아동으로 분류되었을 때 경험할 수 있는 낙인 효과는 이미 어려움을 겪고 있는 아동과 그 가족에게 심각한 결과를 초래할 수도 있다(Dare, 2015). 위험예측모델을 활용해 도움이 필요한 아동과 가족에게 선제적으로 지원을 할 수 있는 반면, 고위험으로 분류된 가족들에 대해 섣부른 판단으로 과잉대응을 할 가능성도 고려되어야 한다. 물론 알고리즘 모델이 선별적 서비스를 제공하는 데 도움이 되는 것은 사실이나 이 접근 방법이 경제적, 사회적으로 취약한 가족들에게 장단기적으로 미칠 수 있는 영향 역시 조심스럽고 면밀하게 연구되어야 한다.

4. 기계학습 편향과 공정성(machine learning bias and fairness)

편향의 위험과 공정성에 대한 이해를 바탕으로 기계학습을 활용하기 위해서는 컴퓨터 과학 분야에서 활발히 논의 중인 기계학습 공정성의 정의와 측정법을 이해하는 것이 도움이 될 수 있다. 컴퓨터 과학에서는 예측 모델의 성능이 인종, 성별, 계급과 같은 개인의 특성에 따라 달라지는지에 초점을 두고 다양한 통계적 접근 방법을 이용해 기계학습 공정성을 평가 및 분석한다(Barocas & Selbst, 2016; Hardt et al., 2016). 기계학습 공정성에 대한 다양한 정의들(Mehrabi et al., 2019) 중 주로 관심을 받은 개념들 중에는 '인지하지 않음을 통

한 공정성'(fairness through unawareness), '개인적 공정성'(individual fairness), '인구 통계적 평등'(demographic parity) 등이 있다. 인지하지 않음을 통한 공정성은 인종, 성별, 계급과 같은 민감한 사회적 특성을 모델링에서 제외시킴으로써(혹은 인지하지 않음으로써) 달성할 수 있다. 개인적 공정성은 비슷한 특성을 가진 개인들이 같은 예측 결과를 받을 때 달성된다. 인구 통계적 평등은 민감한 사회적 특성과 관계없이 모든 사람이 같은 확률로 일정한 예측 결과를 받을 때 이루어진다.

기계학습 공정성에 대한 다양한 접근이 의미 있는 것은 사실이나 최근 일부 학자들은 위와 같이 기계학습 공정성에 통계적으로 접근하는 방법들이 갖는 한계를 지적했다. 통계적 함수를 이용해 좁게 공정성을 정의하면 실제 현실에서 일어나는 복잡한 시스템의 여러 가지 문제를 반영하지 못할 뿐 아니라, 기계학습에 기반한 의사 결정이 개인들에게 끼치는 직간접적 영향도 자주 간과되기 때문이다(Ahmed et al., 2015; McCradden et al., 2020).

5. 나가며

최근 컴퓨터 공학에서 활발하게 논의되는 기계학습 공정성 정의를 아동복지 분야에도 적용하여 알고리즘 위험예측모델의 편향을 분석하려는 시도가 있었으나, 이러한 논의는 아직 기초적인 단계에 머물러 있다(Chouldechova et al., 2018; Coston et al., 2020; Purdy & Glass, 2020). 아동복지 시스템 내 기계학습 활용이 초래할 수 있는 윤리적, 공정성 문제를 최소화하고 아동의 복지와 안전을 최우선하기 위해서는 아동복지 공공 데이터 고유의 특성 및 알고리즘에 기반한 위험예측모델에 대한 깊은 이해와 논의가 필요하다. 이에 더해 시스템을 이용하는 아동과 가족들의 경험 및 의사 결정을 하는 사회복지사들의 관점도 함께 고려되어야 한다.

참고문헌

Ahmed, W., Raza, N., Lodhi, H. W., Muhammad, Z., Jamal, M., & Rehman, A. (2015). Psychosocial factors of antenatal anxiety and depression in Pakistan: Is social support a mediator? *PLoS One*, *10*(1), e0116510. doi: 10.1371/journal.pone.0116510

Ahn, E., Gil, Y., & Putnam-Hornstein, E. (2021). Predicting youth at high risk of aging out of foster care using machine learning methods. *Child Abuse & Neglect*, *117*, 105059. doi: 10.1016/j.chiabu.2021.105059

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, *88*, 402-418. doi: 10.1016/j.eswa.2017.06.035

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2477899

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, *12*(3), 513-544. doi: 10.1111/1745-9133.12047

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, 1-12. doi: 10.1145/3290605.3300271

Brownell, M. D., & Jutte, D. P. (2013). Administrative data linkage as a tool for child maltreatment research. *Child Abuse & Neglect*, *37*(2-3), 120-124. doi: 10.1016/j.chiabu.2012.09.013

Capatosto, K. (2017). Foretelling the future: A critical perspective on the use of predictive analytics in child welfare. *Kirwan Institute Research Report*. The Ohio State University, Kirwan Institute for the Study of Race and Ethnicity. http://kirwaninstitute.osu.edu/wp-content/uploads/2017/05/ki-predictive-analytics.pdf에서 인출.

Chadwick Center and Chapin Hall. (2018). Making the most of predictive analytics: Responsive and innovative uses in child welfare policy and p ractice. *Policy Brief*. Intersection of Research and Policy. https://www.chapinhall.org/wp-content/uploads/Making-the-Most-of-Predictive-Analytics.pdf에서 인출.

Chan, H., Rice, E., Vayanos, P., Tambe, M., & Morton, M. (2017). Evidence from the past: Ai decision aids to improve housing systems for homeless youth. In AAAI 2017 Fall Symposium Series. https://teamcore.seas.harvard.edu/publications/evidence-past-ai-decision-aids-improve-housing-systems-homeless-youth-0에서 인출.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Friedler, S. A. & Wilson, C. (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency*, 81, 134-148). PMLR. http://proceedings.mlr.press/v81/chouldechova18a.html에서 인출.

Coston, A., Mishler, A., Kennedy, E. H., & Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. ArXiv:1909.00066 [Cs, Stat]. http://arxiv.org/abs/1909.00066에서 인출.

Dare, T. (2015). The ethics of predictive risk modeling. In *Challenging child protection*. Jessica Kingsley Publishers.

English, D. J., Brandford, C. C., & Coghlan, L. (2000). Data-based organizational change: The use of administrative data to improve child welfare programs and policy. *Child Welfare*, *79*(5), 499-515.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. ArXiv:1610.02413 [Cs]. http://arxiv.org/abs/1610.02413에서 인출.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction(2nd ed)*. Springer.

Keddell, E. (2015). The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy*, *35*(1), 69-88. doi: 10.1177/0261018314543224

McCradden, M. D., Joshi, S., Mazwi, M., & Anderson, J. A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, *2*(5), e221-e223. doi: 10.1016/S2589-7500(20)30065-0

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ArXiv:1908.09635 [Cs]. http://arxiv.org/abs/1908.09635에서 인출.

Pan, I., Nolan, L. B., Brown, R. R., Khan, R., Van Der Boor, P., Harris, D. G., & Ghani, R. (2017). Machine learning for social services: A study of prenatal case management in Illinois. *American Journal of Public Health*, *107*(6), 938-944. doi: 10.2105/AJPH.2017.303711

Perron, B. E., Victor, B. G., Bushman, G., Moore, A., Ryan, J. P., Lu, A. J., & Piellusch, E. K. (2019). Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child Abuse & Neglect*, *98*, 104180. doi: 10.1016/j.chiabu.2019.104180

Purdy, J., & Glass, B. (2020). The pursuit of algorithmic fairness: On "correcting" algorithmic unfairness in a child welfare reunification success classifier. ArXiv:2010.12089 [Cs, Stat]. http://arxiv.org/abs/2010.12089에서 인출.

Schwartz, I. M., York, P., Nowakowski-Sims, E., & Ramos-Hernandez, A. (2017). Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience. *Children and Youth Services Review*, *81*, 309-320. doi: 10.1016/j.childyouth.2017.08.020

Shroff, R. (2017). Predictive analytics for city agencies: Lessons from children's services. *Big Data*, *5*(3), 189-196. doi: 10.1089/big.2016.0052

U.S. Department of Health & Human Services. (2021). Child maltreatment 2019. Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. https://www.acf.hhs.gov/cb/research-data-technology/ statistics-research/child-maltreatment에서 인출.

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349-391. doi: 10.1007/s10994-015-5528-6

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine*, *45*(3), 354-359. doi: 10.1016/j.amepre.2013.04.022